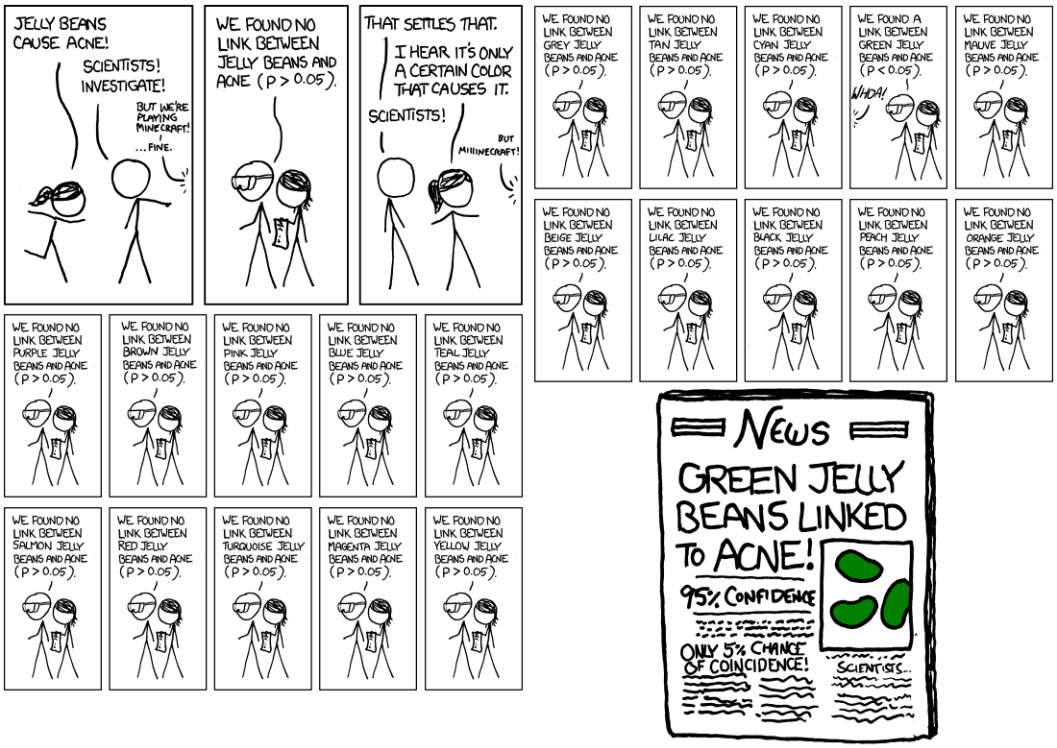
Quantitative Understanding in Biology Short Course
Session 4
Multiple Hypothesis Testing and Non-parametric Tests

Jason Banfelder
Luce Skrabanek

April 1st, 2025



'So, uh, we did the green study again and got no link. It was probably a–'
'RESEARCH CONFLICTED ON GREEN JELLY BEAN/ACNE LINK; MORE STUDY RECOMMENDED!'

http://xkcd.com/882/

1 Multiple Hypothesis Testing

So far, we have considered one statistical test at a time. We know that we can control the rate of Type I errors in these tests by setting α appropriately. In many studies several hypotheses will be tested; for example, we may wish to compare means for several groups (related or independent). We may be testing a drug on several different cell lines, or we may be measuring a response at several different time points after application of a treatment. The proper way to analyze data from studies such as these varies depending on the specifics of the experimental design, and we won't be able to go into detail for all them. There is, however, a common theme that needs to be considered: that of multiple hypothesis testing.

The challenge in dealing with multiple hypothesis testing is controlling the Type I error rate across the whole study. If a study involves testing 20 hypotheses, and each has a 5% chance of a Type I error, then the probability of at least one false conclusion is 64.2%. This result is an application of the binomial distribution where $p = 0.05$ and $n = 20$ ($1 - 0.95^{20} = 0.642$).

In other words, there is a 64% chance that at least one conclusion in our study is wrong. If we want to control the error rate for the study as a whole, we need to adjust α downwards in an appropriate manner. In practice, there are several approaches that are in common use; we'll look at two.

2 The Bonferroni Correction

The simplest and most conservative approach to controlling a study-wide error rate is the Bonferroni correction. Given a desired study-wide error rate, α, you compute a per-test cutoff, α^*, as follows...

$$\alpha^* = \frac{\alpha}{n} \tag{1}$$

...where n is the number of hypothesis tests in your study. In our example above, we compute $\alpha^* = 0.0025$. Using this new per-test cutoff, we can estimate the probability of one or more false conclusions: $1 - (1 - \frac{0.05}{20})^{20} = 0.0488$, which is quite close to what we wanted.

In practice, one usually doesn't adjust α, but rather we 'correct' the p-value. For the Bonferroni correction, we multiply the raw p-value by n to compute the corrected p-value, and compare that to our desired study-wide α of 0.05. While this is a little easier in terms of book-keeping, it should be kept in mind that a 'corrected' p-value is not a probability of any particular scenario we've tested. In fact, corrected p-values can be larger than unity (although they are usually reported as 1 in this case).

The Bonferroni correction is the most conservative correction used in multiple hypothesis

testing. When the number of hypotheses is small, this is probably an appropriate correction to use.

One of the difficulties in critically evaluating scientific literature is that publications are biased toward reporting statistically significant results. When you see a paper that reports a p-value of 0.02 for a particular test, you have no way of knowing how many other hypotheses have been tested and not reported by the authors.

There can also be a problem when you are 'just looking' at some data you have collected to decide how to analyze it. You are implicitly performing many tests on the data, and selecting only those for which the numbers look hopeful to compute a p-value for. In principle, you should be using some kind of multiple hypothesis control in this case.

Ideally, you would design your experiments and your analyses before collecting any data, and all results, statistically significant or not, would be published. As this is not likely or practical in today's scientific and publishing landscape, it is important to recognize that reported results are probably less certain than they might appear, and this may be especially true for retrospective studies.

3 The Benjamini-Hochberg Correction for Controlling False Discovery Rate

The advent of high-throughput biological techniques has resulted in a renewed interest in multiple hypothesis testing. New techniques such as microarray and high-throughput sequencing experiments allow for many thousands of data points to be collected in a single experimental protocol; as a result, it is not uncommon to test tens of thousands of hypotheses in a single analysis. In such cases, many practitioners find that the Bonferroni correction is too conservative.

The most common alternative in use today is the Benjamini-Hochberg correction. It works as follows:

1. Order all raw p-values from smallest to largest, and assign a rank to each one.
2. Correct each p-value by multiplying it by $\frac{n}{\text{rank}}$. This leaves the smallest with the same adjusted p-value as would have been obtained using the Bonferroni correction and the largest p-value uncorrected.
3. Compare the corrected p-values to your pre-determined α, stopping at the first case where the correct p-value is not less than α; all subsequent tests are deemed to be non-significant.

Note that the Benjamini-Hochberg correction seeks to control the "False Discovery Rate", not the "Family-wise Error Rate". This means that we expect some fraction of the signif-

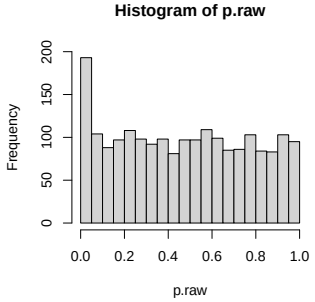
icant result to contain false positives. It is the least conservative correction for multiple hypothesis testing in common use today (short of no correction at all). As such, it is usually applicable to screening studies, where we are trying to identify an enriched set of target genes or compounds for further study, and is usually not the basis on which final scientific conclusions are based.

4 Doing Multiple Hypothesis Testing: A Worked Example

To demonstrate the workings of these multiple hypothesis corrections, we prepared a simulated dataset of a hypothetical screening study. In this study, 2,000 compounds were screened for an effect. For each compound, six controls and six treated samples were assayed, and the results can be compared with a t-test. The raw data are provided as a supplementary file named `screen_raw_data.txt`.

In this dataset, the true positive compounds are the first 100; this is obviously unrealistic in a real-world experiment, but will help us appreciate the impact of the multiple hypothesis correction methods. Additionally, we've synthesized the data such that the signal for the true positive compounds in particularly strong; again, this is a bit unrealistic, but will help illustrate the points here.

You can compute raw p-values for each of the 2,000 compounds by importing the data in Prism as "Grouped" data. Be sure to import and show row titles, so you know to which compound each row refers. Then Analyze with "Multiple t-tests," "One Unpaired t-test per row". If you do this without any multiple-hypothesis testing (and you assume consistent SDs), a plot of the raw p-values will look like the figure below.



We've prepared the histogram with bin widths of 0.05, so the left-most bar contains all of the statistically significant results. The histogram indicates that there are just shy of 200 significant results, which is just about what you'd expect: there are 100 'real' results; plus there should be a Type I error rate of 5%, so we expect $1900 \cdot 0.05 = 95$ false positives.

One important observation to make here is that the distribution of p-values for the true negative cases should follow the uniform distribution (i.e., the histogram should be flat), as we see here. This is generally true, and can be a very useful quality-control metric in studies where you know you have many, many true negatives. It is particularly useful in GWAS studies, where the entire genome is interrogated for effect on a phenotype, and you (more or less) know that only a few regions actually affect the phenotype in a meaningful way.

If we perform a Bonferroni correction (Prism calls this Bonferroni-Dunn method) on our p-values, we see that only 18 values pass the correction. You should understand that we are 95% sure that all 18 of these are true positives; i.e., there is only a 5% chance that there is one or more mistakes in our list. Since we know that the first 100 compounds are the efficacious ones, we can check this by inspecting the row names for each significant result. Indeed, all the compounds listed are true positives. Unfortunately, we've lost many other hits because of the strictness of the Bonferroni correction. And that was with a strong signal!

Let's see how the Benjamini-Hochberg test fares. Do the analysis again, but this time use the "classical" Benjamini-Hochberg correction, and specify a 5% FDR. This analysis should yield 93 hits, or, as Prism terms them "Discoveries". Looking at the row names of the hits, we see that they are all true positives, except for seven. This gives us a False Discovery Rate of $7/93 = 0.075$, which is not far from the expected FDR of 5%. And we've only lost 14 of the 100 truly active compounds.

From the example, you can see why control of FDR is often used in large-scale screening studies. You probably wouldn't want to publish your list of hits from just a Benjamini-Hochberg corrected screen as a definitive list (although too many people do!), but it can be an efficient way to generate high quality leads to be validated by more rigorous (and probably expensive) methods.

5 The Randomization Test: An Example of Testing By Simulation

In a previous session, we saw how to compare two means using the t-test. This test is based on a model in which the data from the two populations are normally distributed and have the same SD. An alternative method for comparing two means, which does not make these assumptions, is called the randomization test. In practice, the randomization test is used rarely, if ever. However, it is interesting because it works without the need for any complex modeling or assumptions. Additionally, the method forms the basis of a non-parametric test for comparing two means, which we will cover shortly.

We begin with two sets of observations, and their means:

$$\begin{array}{ll} x_1, x_2, x_3, x_4, x_5 & \overline{x} \\ y_1, y_2, y_3, y_4, y_5 & \overline{y} \end{array} \tag{2}$$

The difference between the two observed means is

$$\Delta = \overline{y} - \overline{x} \tag{3}$$

As with the t-test, we wish to ascertain whether this observed difference is statistically significant, or if it could be due to chance. Our null hypothesis is that the two sets of observations are samples from the same distribution. Interestingly, for the randomization test we do not need to assume anything about this hypothesized distribution.

Now, if the null hypothesis were true, then any of the values we observed would be just as likely to appear in the first set as in the second. In other words, any rearrangement or shuffling of the values we observed (keeping the count of values in each group the same) is just as likely to have been observed as the arrangement we did in fact observe. We can therefore enumerate every possible rearrangement of the values we observed, and compute a Δ for each one. We then have a histogram of Δs that can serve as an estimate for the probability distribution function of Δ. Using this approximate distribution, we can compute the probability of observing given differences in means from two random samples from our hypothesized distribution. We can then compute what proportion of those Δs is equal to or larger in magnitude than the one we observed. This is the p-value corresponding to our null hypothesis. You can look at the range of Δs, and compute a CI of your choosing.

Of course, this p-value is only an estimate. Interestingly, it is a proportion, so, if you are motivated, you can compute a CI for the p-value using techniques we learned earlier.

As mentioned above, the randomization test is rarely, if ever, used in practice. It involves a good deal of bookkeeping to elencate all of the possible rearrangements of the observed values. For all but the most trivial cases, you would need a computer. Even so, with more than a moderate count of observations in each group, the resultant combinatorial explosion would be beyond the capacity of even the most powerful computers. In such cases, sampling a reasonably large number of rearrangements would allow you to develop an estimate of the distribution of Δ, and would allow you to approximate a p-value. The exhaustive procedure is known as the (ostensibly oxymoronic) exact randomization test!

Again, the randomization test is hardly ever used in practice; most sane people would use the t-test if they were comfortable with its assumptions regarding normality and equivalent SDs. That said, if you are comfortable with the idea behind the randomization test, then you have a good understanding of what a p-value is.

6 The Wilcoxon Rank-Sum Test

The Wilcoxon Rank-Sum test is similar in spirit to the randomization test, and is in fairly common use. It is a non-parametric test that seeks to answer a similar question to the t-test: we again have two sets of observations, and we wish to ascertain whether they come from the same distribution. We are not comfortable with the assumptions of the t-test, and choose a non-parametric method.

Again, we begin with two sets of observations (same as above). We begin our analysis by ordering the values from smallest to largest, and associating a rank with each observation (in the event of a tie, use the average of the ranks to be assigned). Our order might be...

x_4	y_1	y_3	x_2	y_5	y_4	y_2	x_1	x_5	x_3
1	2	3	4	5	6	7	8	9	10

...and finally a Δ for the difference between the rank sums would be...

$$\Delta = 23 - 32 = -9 \tag{4}$$

Our reasoning from this point on is analogous to that of the randomization test. If the samples were from the same underlying distribution, then any rearrangement or shuffling of the data would be just as likely as the arrangement we observed. We can therefore develop a distribution of Δs, and estimate a p-value that indicates how likely we are to see a Δ as large in magnitude as the one we actually observed.

Note that we never used the actual observed values, just their order. As you might imagine, the Wilcoxon Rank-Sum test is quite robust to outliers; it doesn't matter if the largest value is 100 or 10,000,000; the result would be exactly the same.

Note that the test outlined above is sometimes referred to as the "Mann-Whitney test;" this is what Prism calls it. To be very precise, we should call it the "two sample Wilcoxon rank sum test". The 'one sample' version is a non parametric test used for paired data, and is available in Prism under the name "Wilcoxon matched-pairs signed rank test."

7 Homework

1. Analyze the sample data using the Benjamini-Hochberg method with an FDR of 20%. Prepare a table showing the number of TPs, FPs, TNs, and FNs.
2. Read about sensitivity and specificity. Prepare a table showing the sensitivities and specificities of the screen using:
    - raw p-values with a 95% confidence level

- Bonferroni corrected p-values with a 95% confidence level
- Benjamini-Hochberg corrected p-values with a 5% FDR
- Benjamini-Hochberg corrected p-values with a 20% FDR

